Lecture 1
Probability Theory and Statistics

Random Numbers

Distributions

Samples

Confidence Intervals

 Copyright Teemu Kerola 2002

Random Numbers

- Ref: Law-Kelton -91, Chapter 4
- Random number (satunnaisluku)
- Sample space (otosavaruus)
  - discrete **S** = {1, 2, 3, 4, 5, 6}
  - continuous **S** = [0, 1] **S** =[0,1)
- Random variable (satunnaismuuttuja)
  - one way to generate random numbers
  - determined by a function or rule: **R** → **S**
  - random variables: X, Y=5X, Z=g(X,Y)

 Copyright Teemu Kerola 2002

Random Variable X

Prob

- Distribution function $F(x) = P(X \leq x)$
- Density function $f(x) = P(X=x) = P(x)$
- Both F(x) and f(x) determine the (same) distribution
- X discrete?
  - $P(a < X \leq b) = \sum_{a+1}^{b} f(x) = F(b) - F(a)$
- X continuous?
  - $P(a < X \leq b) = \int_a^b f(x) = F(b) - F(a)$

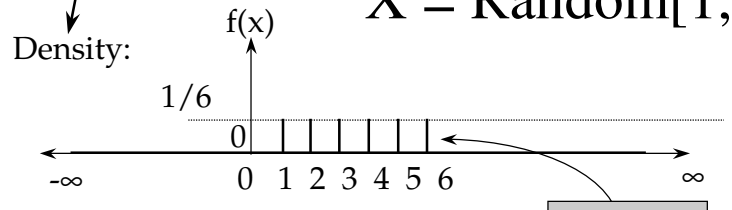
dice: $P(X=6) = 1/6$

temperature: $P(20 \leq X \leq 24) = 80\%$

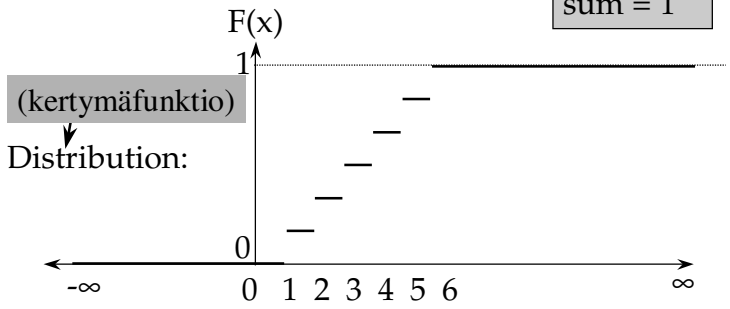
Discrete distribution X = Random[1,6]

(tiheysfunktio)

Density:



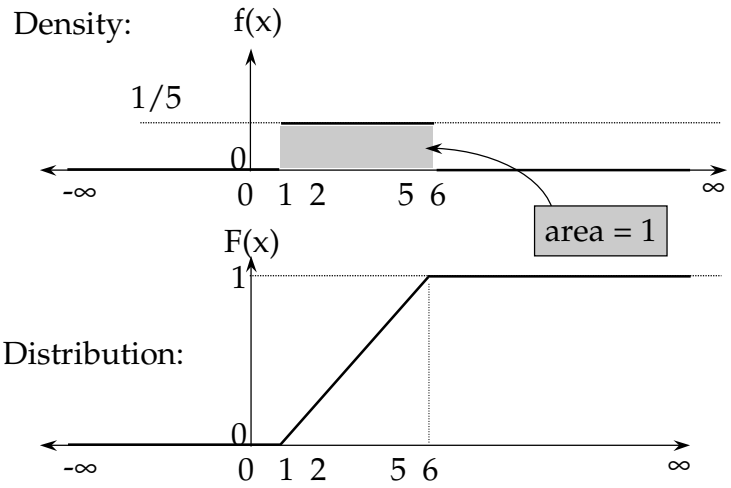
sum = 1



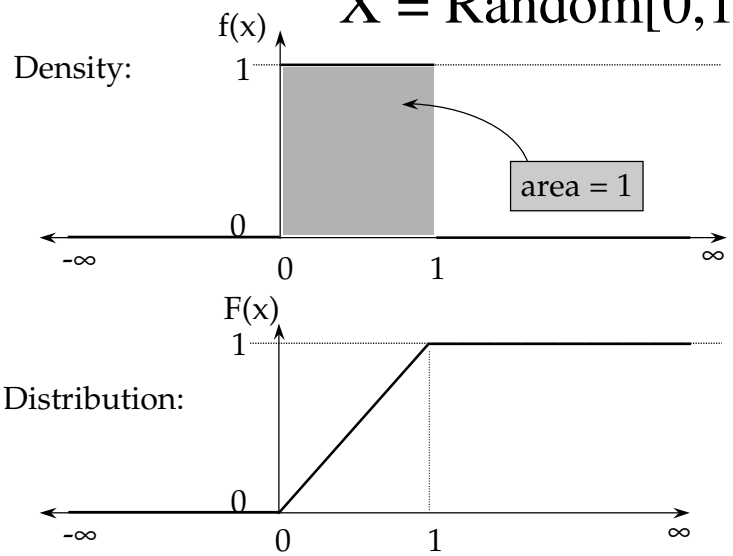
(kertymäfunktio)

Distribution:

Continuous distribution
X = Random[1,6]



Continuous distribution
X = Random[0,1)

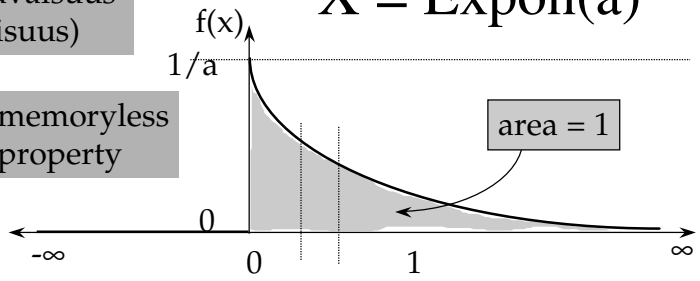


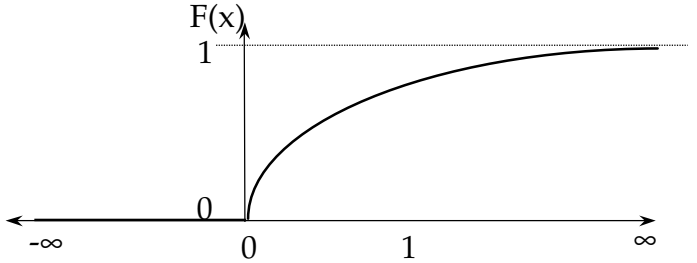
Continuous distribution

X = Expon(a)

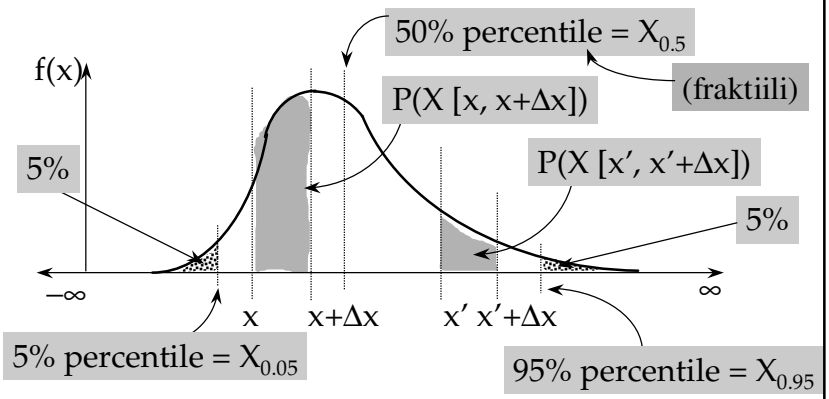
(unohtavaisuus ominaisuus)

memoryless property





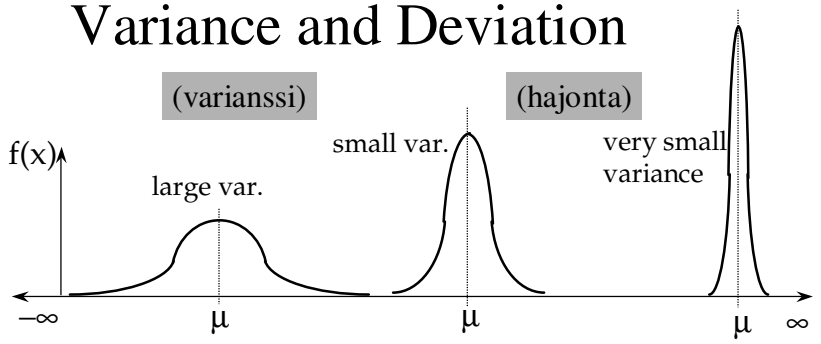
Properties for Random Var. X



(mediaani)

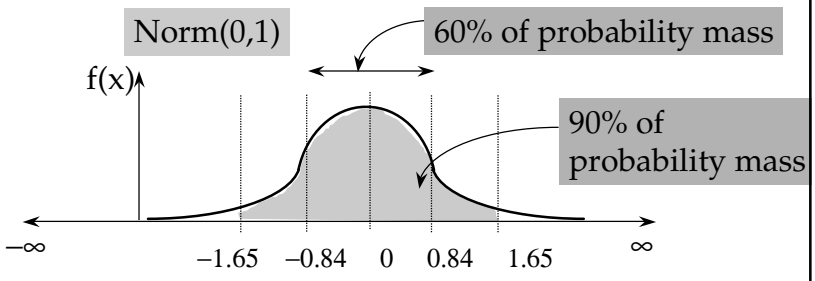
- median $X_{0.5}$: $F(X_{0.5}) = 0.5$
- mean, expected value EX = E(X) = μ (keskiarvo)
  - $EX = \sum_x x\ P(x)$ $\quad EX = \int_x x\ f(x)\ dx$

Variance and Deviation



- average $E(X) = EX = \mu = \int_x x\, f(x)\, dx$
  - describes where the values are located
- variance $Var(X) = \sigma^2 = E[(X-\mu)^2] = E(X^2)-\mu^2$
  - describes how close the values are to average value
- Standard deviation $\sigma = \sqrt{\sigma^2}$

Standard Normal Distribution (2)



- Normal distribution Norm (μ, σ^2)
- Standard Normal Distribution = Z-distribution = Norm (0, 1) = N(0, 1)
- Z-percentiles $Z_{0.95}$ = 1.645 = $Z_{1 - \alpha/2}$ for α = 10%

$Z_{0.80}$ = 0.84

Covariance (kovarianssi)

- X and Y are random variables
  - Does X depend on Y or vice versa or what?
- Covariance $Cov(X,Y) = C_{XY}$
  $= E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X \mu_Y$
- $C_{XY} = C_{YX}$
- $C_{XX} = Cov(X,X) = Var(X) = \sigma^2$
  - $C_{XY} = 0$ X, Y uncorrelated — X large => Y ???
  - $C_{XY} > 0$ X, Y positively correlated — X large => Y large
  - $C_{XY} < 0$ X, Y negatively correlated — X large => Y small

Correlation (korrelaatio)

- Problem: C_{XY} varies much in magnitude
  - can not deduct much from its magnitude
- Solution: scale with (geom. aver) variance
- Correlation $Cor(X,Y) = \rho_{XY} = \frac{C_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}} \in [-1,1]$
  - $\rho_{XY} = 0$ X, Y uncorrelated
  - $\rho_{XY} > 0$ X, Y positively correlated
  - $\rho_{XY} < 0$ X, Y negatively correlated
  - $\rho_{XY} = 0.98$ X, Y highly positively correlated
- Correlation does <u>not</u> mean causal dependence
  - E.g., eating ice cream vs. drowning accidents

Independent X and Y ?

(riippumat-tomuus)

(yhteisjakauman tiheysfuntio)

- Random variables X, Y
- Joint prob. density function $f(X,Y)$:

  $P(X \in A, Y \in B) = \int_A \int_B f(x,y)\, dx\, dy$
- X and Y are independent if

  $f(x,y) = f_X(x) \cdot f_Y(y)$

  where

  $f_X(x) = \int_{-\infty,\infty} f(x,y)\, dy$

  $f_Y(y) = \int_{-\infty,\infty} f(x,y)\, dx$

marginal prob. density functions

Set of Random Variables

- Many random variables Xi or X_i
  - $i \in [1, N]$ $i \in [1, \infty]$
- $E(Xi) = EX_i = \mu_{Xi} = \mu_i$
- $Var(Xi) = Var(X_i) = \sigma_{Xi}^2 = \sigma_i^2$
- $Cov(Xi, Xj) = Cov(X_i,X_j) = C_{Xi,Xj} = C_{i,j} = C_{ij}$
- $Cor(Xi, Xj) = Cor(X_i,X_j) = \rho_{Xi,Xj} = \rho_{i,j} = \rho_{ij}$
- All comparisons pair-wise!

IID Set of Random Variables

- Random variables X_1 , X_2 , ...
- X_i's identically distributed?
  $EX_i = \mu_i = \mu$ $\forall i = 1, 2, ...$
  $Var(X_i) = \sigma_i^2 = \sigma^2$ $\forall i = 1, 2, ...$
  $f_{Xi}(x) = f(x)$ $\forall i = 1, 2, ...$
- X_i's independent?
  X_i and X_j are independent $\forall ij \in \{1, 2, ...\}$
- Now, X_i's are IID
  - independent and identically distributed (= IID)
  - almost never the case in real life
  - almost always needed for the math to work out!

6.3.2002 Copyright Teemu Kerola 2002 15

Sample (otos)

- Problem: What is the distribution of X?
- Solution: try it out and see
- Sample: one set of values for X
  - throw a dice for 100 times? 10000 times?
  - measure response time from 100 trials? Each trial 10 values? 1000 values?
  - write down sales sum for next 10 years?
- New question: what can we tell about X from the sample?
  - What are $EX=\mu$ and $Var(X)=\sigma^2$?

6.3.2002 Copyright Teemu Kerola 2002 16

Sample Properties

- Sample Set $S = \{ X_i \mid i=1,...,n \}$ (otosjoukko)
- Sample points X_i (otospisteet)
  - should be set of IID random variables
  - $X_i \sim X$, i.e., X_i has the (unknown) same distribution than X
    $EX_i=EX=\mu$ and $Var(X_i)=Var(X)=\sigma^2$
- Sample mean $\overline{X}(n) = \frac{\sum_i X_i}{n}$ (otoskeskiarvo)

  (harhaton estimaatti)

  is unbiased estimator for $EX=\mu$: $E\overline{X}(n) = \mu$
  - with large n, sample mean is close to μ
  - how close? how large n?

Sample Properties

- Variance for sample mean: $\text{Var}(\overline{X}(n)) = \frac{\sigma^2}{n}$
  - problem: $\text{var}(X) = \sigma^2$ is unknown
  - solution: use sample variance $s^2(n)$ instead:

$$s^2(n) = \frac{\sum_i [X_i - \overline{X}(n)]^2}{n-1} = \frac{\sum_i X_i^2 - n\overline{X}^2(n)}{n-1}$$

(otosvarianssi)

  - now, $\hat{\text{Var}}(\overline{X}(n)) = \frac{s^2(n)}{n} = \frac{\sum_i [X_i - \overline{X}(n)]^2}{n(n-1)}$

    is unbiased estimator for $\text{Var}(\overline{X}(n))$
  - know now how good estimator sample mean is

Central Limit Theorem (CLT)

- "When n becomes large, normalized sample mean becomes normally distributed"

$$\frac{\overline{X}(n)-\mu}{\sqrt{\sigma^2/n}} \xrightarrow{n\to\infty} N(0,1)$$

(keskeinen raja-arvolause)

- Problem: what if n is not very large?
- Solution: use Student distribution instead:

$$\frac{\overline{X}(n)-\mu}{\sqrt{\sigma^2/n}} \approx T(n-1)$$

(vapausaste)

where n-1 is the degrees of freedom for T

- Percentiles for N(0,1) and T(k) are tabulated

How to Use Central Limit Theorem?

- Use CLT for percentiles

$$P(-z_{1-\alpha/2} \le \frac{\overline{X}(n)-\mu}{\sqrt{\sigma^2/n}} \le z_{1-\alpha/2}) = 1-\alpha$$

- Use unbiased estimator for σ^2

$$P(-z_{1-\alpha/2} \le \frac{\overline{X}(n)-\mu}{\sqrt{s^2(n)/n}} \le z_{1-\alpha/2}) = 1-\alpha$$

- Solve for μ, get confidence interval for μ

$$P(\mu \in \left\{ \overline{X}(n) \pm z_{1-\alpha/2} \sqrt{s^2(n)/n} \right\}) = 1-\alpha$$

(luottamusväli)

Confidence Interval

- (1-α) confidence interval

$$P(\mu \in \left\{ \overline{X}(n) \pm z_{1-\alpha/2} \sqrt{s^2(n)/n} \right\}) = 1-\alpha$$

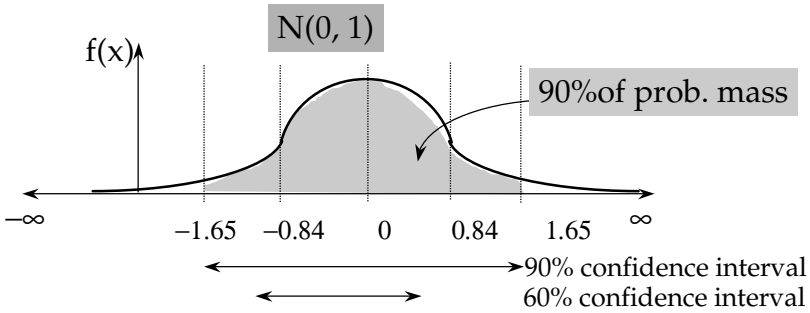
- 90% confidence interval, α = 10%

$$P(\mu \in \left\{ \overline{X}(n) \pm z_{0.95} \sqrt{s^2(n)/n} \right\}) = 90\%$$ n > 30 ?

or

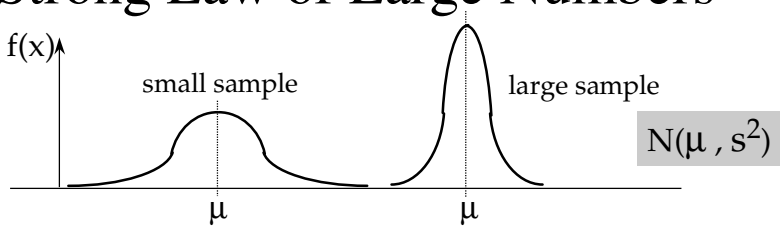
$$P(\mu \in \left\{ \overline{X}(n) \pm t_{n-1,0.95} \sqrt{s^2(n)/n} \right\}) = 90\%$$ n ≤ 30 ?

Confidence Interval (CI)



- Distribution for normalized μ based on sample
- 60% CI is narrower than 90% CI
  - 60% CI looks "better"
  - 60% CI is more likely to be wrong: 40% change that it does not to contain the true mean

Effect of Sample Size, Strong Law of Large Numbers



- sample distribution s^2 for the μ (true, unknown mean) is based on sample
- large sample, s^2 small
- small sample, s^2 large
- Strong Law of Large Numbers: (suurten lukujen laki)

$$\overline{X}(n) \xrightarrow{n\to\infty} \mu \quad \text{with probability 1}$$

Example A

- Question: X is normally distributed. What is the mean μ of X?
  1.20, 1.50, 1.68, 1.89, 0.95, 1.49, 1.58, 1.55, 0.50, 1.09
- Answer:
  - Take sample of 10 observations for X
  - compute $\overline{X}(10)$ and $s^2(10)$ 1.34 and 0.17
  - get 90% percentile for T-distribution with 9 degrees of freedom, $t_{9,\,0.95} = 1.83$
  - compute 90% confidence interval for μ:

$$\overline{X}(10) \pm t_{9,0.95}\sqrt{s^2(10)/10} = 1.34 \pm 1.83*0.13$$
$$= (1.34 \pm 0.24) = (1.10, 1.58)$$

Example A (contd)

- 90% confidence interval for μ:

$$\bar{X}(10) \pm t_{9,0.95}\sqrt{s^2(10)/10} = 1.34 \pm 1.83 * 0.13$$
$$= (1.34 \pm 0.24) = (1.10, 1.58)$$

- There is 90% chance that $1.10 \le \mu \le 1.58$

90% chance that estimate is correct

- There is 10% chance that $\mu < 1.10$ or $1.58 < \mu$

10% chance that estimate is wrong

Example A (contd)

- 80% confidence interval for μ

$$\bar{X}(10) \pm t_{9,0.90}\sqrt{s^2(10)/10} = (1.34 \pm 0.18) = (1.16, 1.52)$$

- 90% confidence interval for μ

$$\bar{X}(10) \pm t_{9,0.95}\sqrt{s^2(10)/10} = (1.34 \pm 0.24) = (1.10, 1.58)$$

- 95% confidence interval for μ

$$\bar{X}(10) \pm t_{9,0.975}\sqrt{s^2(10)/10} = (1.34 \pm 0.29) = (1.05, 1.63)$$

- 99% confidence interval for μ

$$\bar{X}(10) \pm t_{9,0.995}\sqrt{s^2(10)/10} = (1.34 \pm 0.42) = (0.92, 1.76)$$

Example B

- Question: I know the distribution.
  Is the mean $\mu = 5$? ("Two-tailed hypothesis" H_0)
- Answer:
  - Take sample of 10, compute test factor t_9 :

$$t_9 = \frac{\bar{X}(10) - 5}{\sqrt{s^2(10)/10}}$$

(testisuure)

(scaled to std distr)

  - get 90% confidence interval for T-distribution with 9 degrees of freedom $(-t_{9,\,0.95}, t_{9,\,0.95}) = (-1.83, 1.83)$
  - if test factor t_9 is not in there, reject H_0, i.e., say "no"
    - 10% chance of being wrong! (why?)
  - if t_9 is in there, accept H_0, i.e., say "yes"

Example C

- Question: do X and Y have same mean?
- Answer:
  - Take samples for X and Y, sample averages: $\bar{X}(n)$ and $\bar{Y}(m)$
  - get test factor t_{n+m-1} :

$$t_{n+m-1} = \frac{\bar{X}(n) - \bar{Y}(m)}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$$where \;\; s^2 = \frac{1}{n+m-2}\left(\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2\right)$$

  - get 90% confidence interval for T-distribution
  - if t_{n+m-1} is not in there, reject H_0, i.e., say "no"
    - 10% chance of being wrong

Usual Assumptions That Usually Do Not Apply

- Normal distribution
  - central limit theorem
  - strong law of large numbers
- Independent samples
  - X_1, X_2, ...
- Assumptions required by probability theory and statistics to work do not usually apply
- Results may still be usable
- Watch out and always check how far off you are from required assumptions